МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
«ХАРКІВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»

О.Є. Скворчевський

ЕКОНОМЕТРІЯ

Текст лекцій

курсу «Економіко-математичні методи та моделі»

для студентів спеціальностей 6.030601 – «Менеджмент організацій»,
6.030504 – «Економіка підприємства», 6.030507 – «Маркетинг»,
6.030505 – «Управління персоналом та економіка праці»,
6.030507 – «Інтелектуальна власність», 6.030509 – «Облік та аудит»
очної, заочної та дистанційної форм навчання

Затверджено
редакційно-видавничою
радою університету
протокол № __ від _________ 2016 р.

Харків
НТУ «ХПІ»
2016

ВСТУП

Сучасна економіка є значною мірою формалізованою наукою, в якій широко використовуються кількісні методи. Застосування математичних методів в економіці має багато розгалужень і сформувало декілька окремих дисциплін, а саме: дослідження операцій, математичне програмування, теорія ігор тощо. Серед цих дисциплін однією із найважливіших є економетрія, яка входить до циклу базової економічної підготовки не тільки в Україні, але і в Європі, Сполучених Штатах Америки, Російській Федерації та інших провідних країнах. Економетрія є наукою, що в основному застосовує методи математичної статистики для вивчення закономірностей протікання економічних явищ та процесів.

В даному тексті лекцій зроблено спробу, як найвдаліше комбінувати економічний та математичний аспекти економетрії. Конспект лекцій з економетрії розрахований на студентів, що володіють, хоча б в початковому обсязі, основами економічної теорії, мікро- та макроекономіки, лінійної алгебри, математичної статистики. Окрім теоретичних матеріалів в лекціях коротко вказуються найбільш розповсюджені комп’ютерні засоби виконання відповідних розрахунків. Більш повний комп’ютерний інструментарій виконання прикладних математичних досліджень буде представлений у відповідному лабораторному практикумі.

Текст лекцій орієнтовано на студентів, що вперше приступають до вивчення економетрії. Саме тому, акценти робляться на вивченні основних базових тем економетрії та найпростішим прикладним дослідженням в цій галузі. Більша частина конспекту лекції присвячена парному та множинному кореляційно-регресійному аналізу, саме ці теми є ядром економетрії. Студенти мають ґрунтовно вивчити їх перед тим, як перейти до більш складних тем.

Основні матеріали даного конспекту лекцій викладалися семи поколінням студентів економічного факультету Національного технічного університету «Харківський політехнічний інститут». Саме живий плідний діалог зі студентами дозволив автору викласти відносно складні аспекти економетрії в більш-менш доступній та зрозумілій формі. Необхідно виразити подяку тим активним та зацікавленим студентам, які своїми запитаннями, активною участю в науковій роботі, стимулювали автора до корегування та покращення матеріалів, що представлені в роботі.

Незважаючи на значний обсяг теоретичних розробок та прикладних досліджень, економетрія залишається живою та інтенсивно ровиваючоюся наукою. При підготовці, як даного тексту лекцій, так і вцілому навчальних програм з економетрії для студентів економічного факультету Національного технічного університету «Харківський політехнічний інститут» перед автором постала проблема розміщення значного обсягу інформації в дуже обмежену кількість академічних годин, що виділяються на її вивчення студентами загальної економічної підготовки. Вирішення цієї проблеми може бути досягнуто концентрацією на базових темах економетрії, а також вивченню, в основному, прикладних аспектів економетрії. Отже, текст лекцій викладено таким чином, щоб студенти мали уявленні про основні теоретичні положення економетрії, вміли коректно ставити задачу, орієнтувалися в основних комп’ютерних засобах вирішення задач та вміли давати економічну інтерпретацію отриманим результатам.

ТЕМА 1. ЕКОНОМЕТРІЯ. ЇЇ ЗАДАЧІ ТА МЕТОДИ

Сутність економетрії. Типи залежностей між змінними. Емпіричні статистичні данні, як основа економетричних досліджень. Економетрична (регресійна) модель у загальній постановці. Основні етапи і проблеми економетричного моделювання.

1.1. Сутність економетрії

Економетрика, як наука зародилася в перші десятиріччя XX ст., що було пов’язано із усе більшим проникненням математичних методів до різних дисциплін, у тому числі економіки. Протягом усього минулого століття вона стрімко розвивалася. І сьогодні росте кількість наукових публікацій і досліджень із застосуванням економетричних методів. Свідченням високого світового визнання економетрії є присудження за найбільш значні розробки у цій області Нобелівських премій по економіці Р.Фрішу і Я. Тінбергу (1969), Л. Клейну (1980), Т. Хаавельмо (1989), Дж. Хекману і Д.Макфаддену (2000).

В нашій країні у період централізованої планової економіки упор робився на балансових та оптимізаційних методах дослідження, на описанні «системи функціонування соціалістичної економіки», побудові оптимізаційних моделей галузей та підприємств, то при ринковій економіці підвищується роль економетричних методів. Без знання цих методів неможливо ні дослідження та теоретичне узагальнення емпіричних залежностей економетричних змінних, ні побудування надійного прогнозу в банківській справі, фінансах та бізнесі [1].

Єдине загальноприйняте визначення економетрії в даний час відсутнє. На думку авторів найбільш точним та повним є класичне визначення економетрії дане С. Фішером:

Економетрія – це розділ економіки, що займається розробкою та застосуванням статистичних методів для виміру взаємозв’язків між економічними змінними.

У літературі також зустрічається інша назва цієї дисципліни – економетрика.

Основною *метою економетрії* є аналіз та математичний опис спостерігаємих економічних явищ та процесів, як правило, із метою прогнозування розвитку цих процесів для прийняття науково-обґрунтованих управлінських рішень на базі прогнозів.

Із вищесказаного можна зробити висновок, що *предметом економетрії* є дескриптивні економіко-математичні моделі.

Об'єктом економетрії можуть бути різні економічні системи від окремих підприємств до глобальної економіки вцілому.

1.2 Типи залежностей між змінними

Важливим етапом вивчення взаємозв'язку між змінними є встановлення типу залежностей між ними. Якщо кожному значенню однієї змінної відповідає конкретне значення іншої змінної, то між ними існує *функціональна залежність*. Такий тип залежностей часто зустрічається у фізиці та техніці і значно рідше в економіці. Серед найпростіших прикладів функціональних залежностей в економіці можна навести залежність вартості покупки віт кількості одиниць товару та ціни, обсяг податкових виплат від ставки податку тощо.

В економіці дуже часто значення змінних величин носить випадковий характер. Залежність між двома випадковими величинами називається *імовірнісною* (стохастичною або статистичною), якщо кожному значенню однієї з них відповідає певний діапазон розподілення іншої. Серед таких залежностей можна назвати наступні: залежність попиту від ціни, залежність витрат підприємства від кількості виробленої продукції, залежність відсотку бракованої продукції від відсотку робітників зі спеціальною підготовкою тощо.

Виділяють також *регресійну залежність* – однобічну залежність випадкової змінної від однієї або декількох невипадкових змінних [1].

1.3. Емпіричні статистичні данні, як основа економетричних досліджень

Економетричні моделі будуються, як правило, на основі емпіричних статистичних даних. Для того, щоб отримати достовірні оцінки параметрів моделі необхідно мати випадкову вибірку спостережень достатньо великого об'єму.

Послідовність спостережень $x_1,...,x_n$ називається *випадковою вибіркою об'єму* n, якщо $x_1,...,x_n$ отримані, як незалежні реалізації деякої випадкової величини x із розподіленням $F(x)$. При цьому також кажуть, що $x_1,...,x_n$ є вибірка із генеральної сукупності x чи $F(x)$. Із теоретико-імовірнісної точки зору випадкова вибірка $x_1,...,x_n$ може розглядатися, як послідовність незалежних випадкових величин, що мають одне і теж розподілення $F(x)$ [3].

В класичному курсі економетрії розглядається два типа вибіркових даних: просторова вибірка та часовий ряд.

У економіці під *просторовою вибіркою* (просторовими даними) розуміють набір показників економічних змінних одержаний в даний момент часу. Проте, таке визначення не дуже зручне, через неоднозначність поняття «момент часу». Це може бути і день, і тиждень, і рік. Очевидно, про просторову вибірку має сенс говорити в тому випадку, якщо всі спостереження одержані приблизно в незмінних умовах, тобто є набором незалежних вибіркових даних з деякої генеральної сукупності. Прикладом просторових даних є, наприклад, набір відомостей (об'єм виробництва, кількість працівників, дохід і ін.) по різних фірмах в один і той же момент часу (просторовий зріз). Іншим прикладом можуть бути дані по курсах купівлі/продажу готівкової валюти в один і той же день по обмінних пунктах в Харкові.

Часовим (динамічним) рядом називається вибірка спостережень, в якій важливі не тільки самі значення випадкових величин, але і порядок їх проходження ода за одною. Найчастіше впорядкованість обумовлена тим, що експериментальні дані є серією спостережень однієї і тієї ж випадкової величини в послідовні моменти часу. В цьому випадку динамічний ряд називається часовим рядом. При цьому передбачається, що тип розподілу спостерігаємої випадкової величини залишається одним і тим же (наприклад, нормальним), але параметри його міняються залежно від часу.

Моделі часових рядів, як правило, виявляються складнішими за моделі просторової вибірки, оскільки спостереження у разі тимчасового ряду взагалі кажучи не є незалежними, а це значить, що помилки регресії можуть корелювати один з одним, що значно ускладнює статистичний аналіз моделі. Прикладами тимчасових даних можуть бути щоквартальні дані по інфляції, середній заробітній платні, національному доходу, грошовій емісії за останні роки або, наприклад, щоденний курс валют, встановлений НБУ.

Проте, маючи тільки ряд спостережень без розуміння їх природи, неможливо визначити, маємо ми справу з просторовою вибіркою або тимчасовим рядом [1, 3].

1.4. Економетрична (регресійна) модель у загальній постановці

Економетрична модель – вірогідністно-статистична модель, що описує спостерігаєме економічне явище чи процес та може давати можливість прогнозування їх розвитку.

У загальному випадку економетричну модель можна записати у вигляді:

$$y = \beta_0 + \beta_1 \cdot f(x_1,\ldots,x_m) + \varepsilon , \tag{1.1}$$

де y – залежна змінна;

$x_1,\ldots,x_m$ – незалежні змінні;

m – кількість незалежних змінних;

β_0, β_1 – параметри;

ε – похибка, спричинена випадковими процесами;

Економетрична модель (1.1) містить параметри та змінні, які поділяються на залежні та незалежні.

Параметри – коефіцієнти, що визначаються. як правило, на основі статистичних даних та для усієї вибірки спостережень приймають або фіксоване значення, або лежать в межах деякого інтервалу значень.

Незалежні змінні – змінні, що визначаються поза моделлю, автономно від неї, керовані та плануємі [2].

Незалежні змінні також називаються *екзогенними*, пояснюючими, вхідними тощо або факторами.

Залежні змінні – змінні, значення котрих формуються в процесі функціонування економічної системи, яку описує модель, у значній мірі під впливом незалежних змінних, а також у взаємодії одна із одною [2].

Незалежні змінні також називаються *ендогенними*, поясненими, вихідними тощо або показниками.

Із точки зору математичної статистики формула (1.1) є моделлю регресії. Необхідно відмітити, що економетричні моделі не завжди є регресійними моделями, наприклад через наявність систематичних похибок [1-5].

1.5. Основні етапи і проблеми економетричного моделювання

Можна виділити шість основних етапів економетричного моделювання постановочний, апріорний, етап параметризації, інформаційний, етапи ідентифікації і верифікації моделі [1, 4].

Зупинимося докладніше на кожному з цих етапів і розглянемо проблеми, пов'язані з їх реалізацією.

1-й етап (постановочний). Формується мета дослідження, набір економічних змінних, що беруть участь в моделі.

Як мета економетричного моделювання звичайно розглядається аналіз досліджуваного економічного об'єкту (процесу); прогноз його економічних показників, імітацію розвитку об'єкту при різних значеннях екзогенних змінних (відображаючи їх випадковий характер, зміна в часі), вироблення управлінських рішень.

При виборі економічних змінних необхідне теоретичне обґрунтування кожної змінної (при цьому рекомендується, щоб число їх було не дуже великим і, як мінімум, у декілька разів менше числа спостережень). Незалежні змінні не повинні бути зв'язані функціональною або тісною кореляційною залежністю, оскільки це може привести до неможливості оцінки параметрів моделі або до отримання нестійких, які не мають реального сенсу оцінок, тобто до явища мультіколлінеарності.

Забігаючи вперед, відзначимо, що для відбору змінних можуть бути використані різні методи, зокрема процедури покрокового відбору змінних. А для оцінки впливу якісних ознак (наприклад, стать, освіта і т. п.) можуть бути використані фіктивні змінні. Але що у будь-якому випадку визначає при включенні в модель тих або інших змінних є економічний (якісний) аналіз об'єкту.

2-й етап (апріорний). Проводиться аналіз суть об'єкту, що вивчається, формування і формалізація апріорної (відомої на початку моделювання) інформації.

3-й етап (параметризація). Здійснюється безпосередньо моделювання, тобто вибір загального виду моделі, виявлення зв'язків між змінними, що входять до неї.

Основне завдання, що вирішується на цьому етапі, – вибір виду функції в економетричній моделі, зокрема, можливість використання лінійної моделі як найбільш простій і надійної. Вельми важливою проблемою на цьому (і попередніх) етапі економетричного моделювання є проблема специфікації моделі, зокрема: вираз в математичній формі виявлених зв'язків і співвідношень; встановлення складу незалежних і залежних змінних, формулювання початкових передумов і обмежень моделі. Від того, наскільки вдало вирішена проблема специфікації моделі, в значній мірі залежить успіх всього економетричного моделювання.

4-й етап (інформаційний) Здійснюється збір необхідної статистичної інформації – спостерігаємих значень економічних змінних:

$$X_1 = (x_{i1}, x_{i2}, \dots, x_{in}), \tag{1.2}$$

$$Y = (y_{i1}, y_{i2}, \dots, y_{in}), \tag{1.3}$$

де $i = 1, \dots, n$ кількість спостережень.

Тут можуть бути спостереження, одержані як за участю дослідника, так і без його участі (в умовах активного або пасивного експерименту).

5-й етап (ідентифікація моделі). Здійснюється статистичний аналіз моделі і оцінка її параметрів. Реалізації цього етапу присвячена основна частина підручника.

З проблемою ідентифікації моделі не слід плутати проблему її ідентіфікуємості, тобто проблему можливості отримання однозначно певних параметрів моделі, заданою системою одночасних рівнянь.

6-й етап (верифікація моделі). Проводиться перевірка істинності, адекватності моделі. З'ясовується, наскільки вдало вирішені проблеми специфікації, ідентифікації і ідентіфікуємості моделі, яка точність розрахунків по даній моделі, кінець кінцем, наскільки відповідає побудована модель модельованому реальному економічному об'єкту або процесу.

Приведене вище розділення економетричного моделювання на окремі етапи носить в значній мірі умовний характер, оскільки ці етапи можуть перетинатися, взаємно доповнювати один одного і т.д.

Запитання для самоперевірки

1. Дати визначення та розкрити сутність економетрії.
2. Назвати та дати визначення основним типам залежностей між змінними.
3. Емпіричні статистичні дані, як основа економетричних досліджень.
4. Дати визначення просторової вибірки та часового ряду.
5. Назвати принципову відмінність між ними.
6. Економетрична (регресійна) модель в загальній постановці та її складові.
7. Назвати основні етапи і проблеми економетричного моделювання.

Список літератури

1. Кремер Н.Ш., Прутко Б.А. Эконометрика: Учебник для вузов / Под. ред. проф. Н.Ш. Кремера. – М.: ЮНИТИ-ДАНА, 2007. – 311 с.
2. Эконометрика / Сост. Леванова Л.Н. – Саратов.: Российский государственный технический университет путей сообщения, 2003. – 34 с.
3. Магнус Я.Р., Катышев П.К., Пересецкий А.А. Эконометрика. Начальный курс: Учеб. – 6-е изд., перераб. и доп. – М.: Дело, 2004. – 576 с.
4. Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрики. – М.:ЮНИТИ, 1998. – 310 с.
5. Эконометрика: Учебник / Под ред. И.И. Елисеевой. – М.: Финансы и статистика, 2004. – 344 с.

ТЕМА 2. ОЦІНКА ТІСНОТИ ЛІНІЙНОГО ЗВ'ЯЗКУ МІЖ ДВОМА ВИПАДКОВИМИ ВЕЛИЧИНАМИ

Парний лінійний коефіцієнт кореляції. Оцінка тісноти та напрямку лінійного зв'язку за парним коефіцієнтом кореляції. Кореляційні поля.. Перевірка статистичної значущості коефіцієнта кореляції. Комп'ютерні засоби розрахунку парного лінійного коефіцієнту кореляції та перевірки його статистичної значущості. Коваріація.

Як було сказано вище, основою економетричних досліджень, як правило, є величини представлені емпіричними статистичними даними, наприклад (2.1), (2.2). Перед проведенням ґрунтовних досліджень взаємозв'язків таких величин доцільно попередньо оцінити тісноту зв'язку між ними. Саме цим питанням і присвячена дана тема.

Парний коефіцієнт кореляції r – міра тісноти *лінійного* зв'язку двох випадкових величин X та Y, поданих у вигляді двох вибірок однакового обсягу n:

$$X = \begin{pmatrix} x_1 \\ \dots \\ x_i \\ \dots \\ x_n \end{pmatrix}, \tag{2.1}$$

$$Y = \begin{pmatrix} y_1 \\ \dots \\ y_i \\ \dots \\ y_n \end{pmatrix}. \tag{2.2}$$

Парний коефіцієнт кореляції часто називають вибірковим або коефіцієнтом кореляції К. Пірсона. Як повну назву величини r можна використовувати таку: вибірковий лінійний парний коефіцієнт кореляції К. Пірсона, однак у літературі, як правило, користуються більш короткими назвами [1–6]. Парний коефіцієнт кореляції може бути розрахований за формулою [1, 2]:

$$r = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\sqrt{(\overline{x^2} - \overline{x}^2) \cdot (\overline{y^2} - \overline{y}^2)}}, \tag{2.3}$$

де $\overline{x}$, $\overline{y}$ – середнє значення незалежної та залежної змінних відповідно;

$\overline{xy}$ – середнє значення добутку незалежної та залежної змінних;

$\overline{x^2}$, $\overline{y^2}$ – середнє значення квадратів незалежної та залежної змінних відповідно.

Відповідні середні визначаються за формулами:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}; \tag{2.4}$$

$$\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}; \tag{2.5}$$

$$\overline{xy} = \frac{\sum_{i=1}^{n} (x_i \cdot y_i)}{n}; \tag{2.6}$$

$$\overline{x^2} = \frac{\sum_{i=1}^{n} x_i^2}{n}. \tag{2.7}$$

Коефіцієнт кореляції може набувати значення від –1 до 1. За знаком та абсолютним значенням коефіцієнта кореляції можна зробити такі висновки:

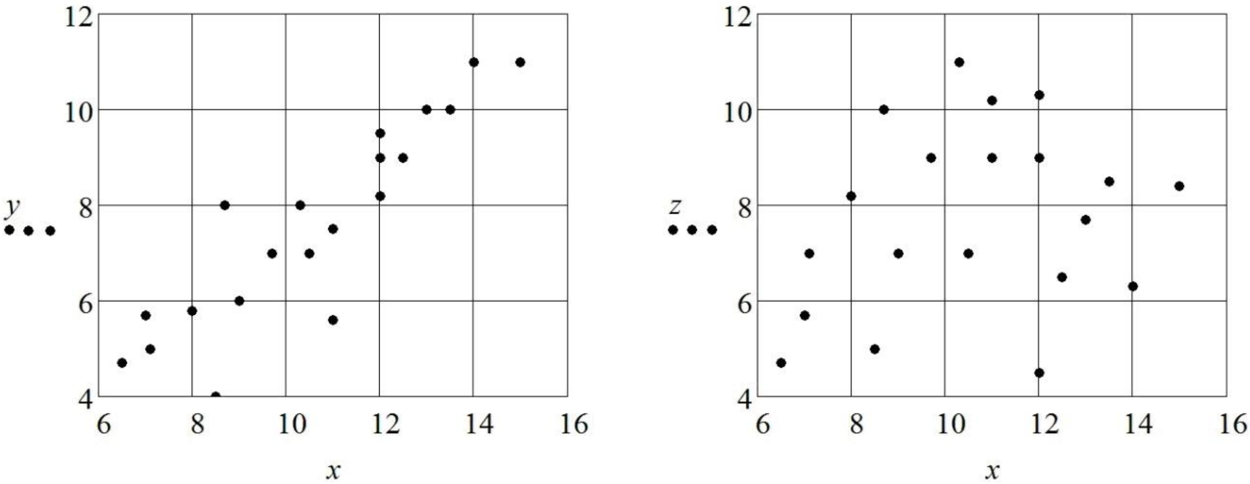
1. Чим ближче абсолютне значення $|r|$ коефіцієнта кореляції до 1, тим тісніший зв'язок існує між двома змінними x та y (рис. 2.1 *а*), а чим ближче $|r|$ до нуля тим зв'язок слабкіший (рис. 2.1 *б*). Якщо парний коефіцієнт кореляції

дорівнює 1, то між двома змінними існує лінійна функціональна залежність. При цьому усі спостереження на кореляційному полі розташовуються вздовж прямої лінії. Якщо парний коефіцієнт кореляції дорівнює 0, то між двома змінними залежність відсутня.

Для перетворення кількісної характеристики тісноти лінійного зв’язку між двома випадковими у якісну може бути використана шкала англійського статистика Чеддока (табл. 2.1).

Таблиця 2.1 – Шкала Чеддока

Величина абсолютного значення парного коефіцієнта кореляції	Характеристика лінійного зв’язку між двома випадковими величинами
до 0,3	практично відсутній
0,31–0,5	слабкий
0,51–0,7	помітний
0,71–0,9	сильний
0,91–0,99	дуже сильний



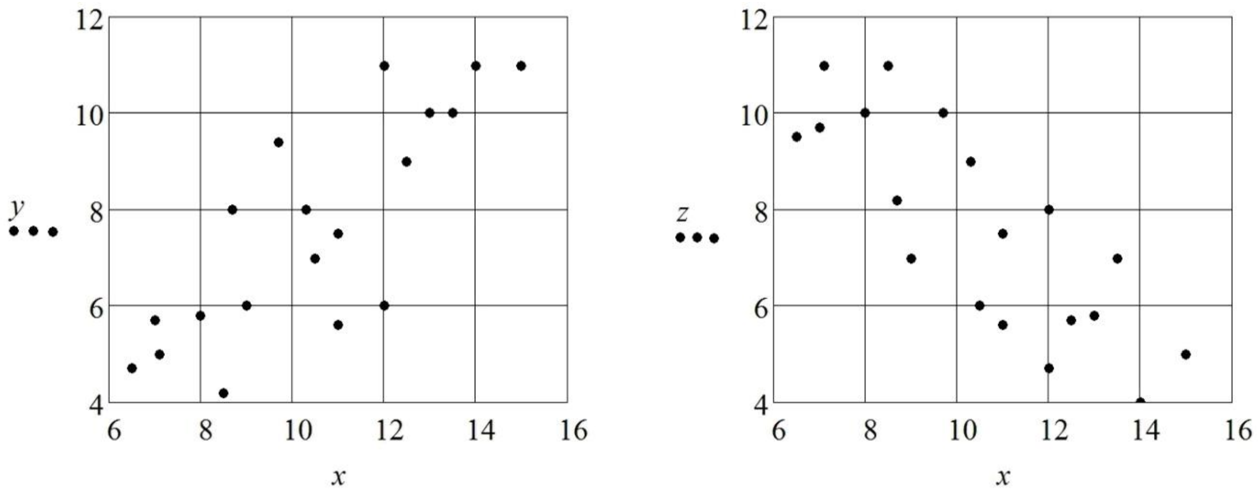
$corr(x, y) = 0{,}898$ $\qquad$ $corr(x, z) = 0{,}242$

а $\qquad$ *б*

Рисунок 2.1 – Кореляційні поля: *а* – для вибірки сильнокорельованих даних; *б* – для вибірки слабокорельованих даних

2. Якщо коефіцієнт кореляції позитивний $r > 0$, то зв'язок між змінними прямий, тобто при збільшенні незалежної змінної x залежна змінна y теж

збільшується (рис. 2.2 *а*). При $r < 0$ зв'язок між змінними зворотний, тобто при збільшенні незалежної змінної x залежна змінна y зменшується (рис. 2.2 *б*).



$corr(x, y) = 0{,}8$ $corr(x, z) = -0{,}8$

а *б*

Рисунок 2.2 – Кореляційні поля: *а* – для вибірки даних з прямим зв’язком; *б* – для вибірки даних зі зворотним зв’язком

Парний коефіцієнт кореляції є випадковою величиною, оскільки обчислюється для випадкових величин. Для нього необхідно висувати і перевіряти гіпотезу про те, чи статистично значуще він відрізняється від нуля (тобто чи є взаємозв'язки між величинами). Ця гіпотеза перевіряється за допомогою *t*-критерію Ст’юдента, фактичне значення якого визначається за формулою:

$$t_{\text{факт}} = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}}. \quad (2.8)$$

Фактичне значення критерію Ст’юдента порівнюється із критичним $t_{p;k}$, отриманим за відповідною таблицею (додаток А) для степенів свободи $k = n - 2$ та вірогідності p, яка в загальному випадку може обиратися довільно. В багатьох джерелах як критерій вибору пропонується рівень значущості $\alpha = 1 - p$. На сучасному етапі розвитку економетрії та інших наук, пов’язаних із математичною статистикою, критичне значення критерію Ст’юдента можна визначити за допомогою функції СТЬЮДРАСПОБР Microsoft Excel 2003-2007 та функції СТЬЮДЕНТ.ОБР.2Х Microsoft Excel 2010. При використанні цих функцій у рядку «Вірогідність» діалогового вікна потрібно зазначати не вірогідність *p*, а рівень значущості α. Така невідповідність, скоріше за все, пов’язана із неточністю перекладу інтерфейсу статистичних функції із англійської мови на російську. В MathCAD критичне значення *t*-критерію Ст’юдента можна отримати за допомогою оператора $qt(p,k)$.

Якщо $|t_{\text{факт}}| > t_{p;k}$, то коефіцієнт кореляції статистично значуще відрізняється від нуля та залежність є достовірною. У протилежому випадку коефіцієнт кореляції статистично не значущий та кореляційний зв'язок між змінними відсутній.

Очевидно, що різним рівням вірогідності будуть відповідати різні значення $t_{p;k}$. Тобто при меншій вірогідності коефіцієнт кореляції може бути статистично значущим, а при більшій – не значущим. Тому обов’язково необхідно вказувати при якій вірогідності було обрано *t*-критерій Ст’юдента. Це можна зробити, відзначивши, що *коефіцієнт кореляції статистично значущий (або не значущий) при вірогідності, наприклад,* 0,9. Вказати вірогідність p (чи рівень значущості α) та ступінь свободи, при якому обиралося табличне значення *t*-критерію можна в його індексі – $t_{p;k}$.

Зі збільшенням вірогідності p критичне значення *t*-критерію Ст’юдента буде зростати. Тому, якщо парний коефіцієнт кореляції виявився статистично не значущим для високої вірогідності (наприклад 0,95–0,99), доцільно перевірити його статистичну значущість при меншому рівні вірогідності (наприклад 0,8–0,85).

При проведенні кореляційного аналізу потрібно пам’ятати, що парний коефіцієнт кореляції є мірою тісноти *лінійного* зв’язку між двома випадковими величинами. Тому мале абсолютне значення коефіцієнта кореляції $|r|$ і (або) його статистична незначущість може свідчити лише про відсутність чи слабкість *лінійного* зв’язку, в той час як між двома випадковими змінними *може існувати тісний нелінійний зв'язок*. Одним із способів виявити такий зв’язок є аналіз відповідних кореляційних полів (див. тему «Нелінійна регресія»).

Коефіцієнт кореляції лише констатує факт присутності чи відсутності лінійного зв’язку між двома величинами X та Y, але він не вказує на причино-наслідкові зв’язки між ними. Таким чином визначити яка змінна є причиною, а яка наслідком можна, виходячи із економічного змісту задачі. Також в економіці зустрічаються випадки, коли тісний кореляційний зв'язок між двома величинами X та Y пояснюється дією третього фактора Z, що одночасно діє і на X, і на Y. Причому причинно-наслідкові зв’язки безпосередньо між X та Y відсутні.

Формула (2.3) оцінки коефіцієнта кореляції рекомендується до застосування при великій кількості спостережень та якщо r не близьке до $|1|$. Якщо величина коефіцієнта кореляції близька до 1, то розподіл його оцінок відрізняється від нормального або розподілу Ст`юдента, оскільки величина коефіцієнта кореляції обмежена значеннями від – 1 до + 1. Щоб обійти це ускладнення для оцінки істотності коефіцієнта кореляції вводиться допоміжна величина z [2].

Математичні вирази, що знаходяться в чисельнику та знаменнику формули (2.3), мають самостійний статистичний зміст. Так величина

$$Cov(X,Y) = \overline{xy} - \overline{x} \cdot \overline{y} \tag{2.9}$$

називається *вибірковим кореляційним моментом* або *вибірковою коваріацією.* Ця величина є мірою тісноти лінійного зв’язку двох випадкових величин X та Y, однак, на відміну від коефіцієнта кореляції вибіркова коваріація має розмірність, що ускладнює її практичне застосування [1, 3]. Знак вибіркової коваріації має таку ж інтерпретацію, як і знак коефіцієнта кореляції.

У знаменнику формули (2.3) знаходиться корінь квадратний із добутку вибіркових дисперсій двох випадкових величин X та Y:

$$Var(X) = \overline{x^2} - \overline{x}^2; \tag{2.10}$$

$$Var(Y) = \overline{y^2} - \overline{y}^2. \tag{2.11}$$

Вибіркова дисперсія (*варіація*) – оцінка дисперсії випадкової величини за вибіркою. Під дисперсією випадкової величини тут потрібно розуміти міру розсіювання випадкової величини, тобто її відхилення від математичного очікування. Вибіркова дисперсія X позначається, як $Var(X)$ від англ. variance, також можуть бути такі позначення $D[x]$, σ_x^2. Дисперсія $Var(X)$ вимірюється у квадраті одиниці вимірювання випадкової величини X, що є незручним. Квадратний корінь із дисперсії називається *середньоквадратичним відхиленням* або *стандартним відхиленням.* Стандартне відхилення вимірюється у тих же величинах, що і сама випадкова величина.

Враховуючи формули (2.9)–(2.11), коефіцієнт кореляції (2.3) може бути розрахований так:

$$r = \frac{Cov(X,Y)}{\sqrt{Var(X) \cdot Var(Y)}}. \qquad (2.12)$$

На сучасному етапі розвитку наук, пов'язаних із математичною статистикою, зокрема економетрії, немає необхідності здійснювати громіздкі розрахунки за формулами (2.3) чи (2.12). У Microsoft Excel існує функція «КОРРЕЛ», яка дозволяє розрахувати парний коефіцієнт кореляції на основі двох вибірок статистичних даних, а в MathCAD – функція *corr(X,Y)*. Таким чином, досліднику залишається лише коректно ввести вихідні статистичні данні в комп'ютерну програму та зробити правильні висновки за отриманими результатами.

В даній темі розкриваються ті питання економетричних досліджень, що дозволяють лише оцінити тісноту лінійного зв'язку між двома величинами, представленими статистичними даними. Якщо, встановлено, що між величинами існує тісний кореляційний зв'язок та відслідковується причинно-наслідковий зв'язок, то можна перейти до визначення аналітичної залежності, що буде його характеризувати.

Запитання для самоперевірки

1. Дати визначення парного лінійного коефіцієнта кореляції.
2. Розрахунок парного лінійного коефіцієнта кореляції.
3. В яких межах може змінюватись парний лінійний коефіцієнт кореляції?

4. Які висновки можна зробити за значенням парного лінійного коефіцієнта кореляції?

5. Розкрити відповідність значень коефіцієнта кореляції та вигляду відповідних кореляційних полів.

6. Порядок перевірки статистичної значущості коефіцієнта кореляції.

7. Чи визначає коефіцієнт кореляції причино-наслідкові зв'язки між величинами представленими вибірками статистичних даних?

8. Чи завжди мале значення коефіцієнта кореляції свідчить про відсутність будь-якого зв'язку між змінними?

9. Коротко розкрити поняття коваріації.

10. Комп'ютерні засоби розрахунку парного лінійного коефіцієнту кореляції та перевірки його статистичної значущості.

Список літератури

1. Кремер Н.Ш. Эконометрика : учебник для вузов / Н.Ш. Кремер, Б.А. Путко ; под ред. Н.Ш. Кремера. – М. : ЮНИТИ-ДАНА, 2007. – 311 с.

2. Елисеева И.И. Эконометрика : учеб. / И.И. Елисеева ; под ред. И.И. Елисеевой. – М. : Финансы и статистика, 2004. – 344 с.

3. Доугерти К. Введение в эконометрику / К. Доугерти ; пер. с англ. – М. : ИНФРА-М, 1999. – 402 с.

4. Магнус Я.Р. Эконометрика. Начальный курс : учеб. / Я.Р. Магнус, П.К. Катышев, А.А. Пересецкий. – 6-е изд., перераб. и доп. – М. : Дело, 2004. – 576 с.

5. Айвазян С.А. Прикладная статистика и основы эконометрики / С.А. Айвазян, В.С. Мхитарян. – М. : ЮНИТИ, 1998. – 432 с.

6. Джонстон Дж. Эконометрические методы / Дж. Джонстон ; пер. с англ. – М.: Статистика, 1980. – 444 с.

7. Лещинський О.Л. Економетрія : навч. посіб. для студ. вищ. навч. закл. / О.Л. Лещинський, В.В. Рязанцева, О.О. Юнькова. – К. : МАУП, 2000. – 208 с.

8. Кремер Н.Ш. Математическая статистика. – М.: Економическое образование, 1992. – 203 с.

ТЕМА 3. ПАРНА ЛІНІЙНА РЕГРЕСІЯ

Сутність парного лінійного регресійного аналізу. Метод найменших квадратів. Альтернати методу найменших квадратів. Теорема Гауса-Маркова. Економічний зміст параметрів парної лінійної регресії.

3.2. Сутність парного лінійного регресійного аналізу

Регресійний аналіз – статистичний метод дослідження залежності випадкової змінної Y від однієї X або декількох незалежних змінних $X_1,...,X_m$. Завданнями регресійного аналізу є встановлення форми залежності між змінними, оцінка функції регресії, прогноз значень залежної змінної. Будь-яке економетричне дослідження починається зі *специфікації моделі*, тобто із вибору структури моделі та визначення набору пояснюючих змінних. Вибір структури моделі може здійснюватися на основі економічної теорії, попередніх результатів, апріорних знань або спеціальних математичних процедур. Парна регресія достатня, якщо існує домінуючий фактор, який і використовується як незалежна змінна. Оскільки найбільш простою формою залежності в математиці є лінійна, то в регресійному аналізі найбільш популярні лінійні моделі. Саме їм приділяється основна увага у базовому курсі економетрії. Таким чином, у рамках цього пункту розглянемо парний лінійний регресійний аналіз, тобто статистичний метод дослідження лінійної залежності випадкової змінної Y від однієї X незалежної змінної. Тісноту лінійного зв’язку між ними можна визначити за парним коефіцієнтом кореляції, а причинно-наслідкові зв’язки – виходячи із економічного змісту задачі.

Парна лінійна регресія – причинна модель статистичного лінійного зв'язку між двома кількісними змінними Y та X, подана рівнянням:

$$\hat{Y} = \beta_0 + \beta_1 \cdot X + \varepsilon, \tag{3.1}$$

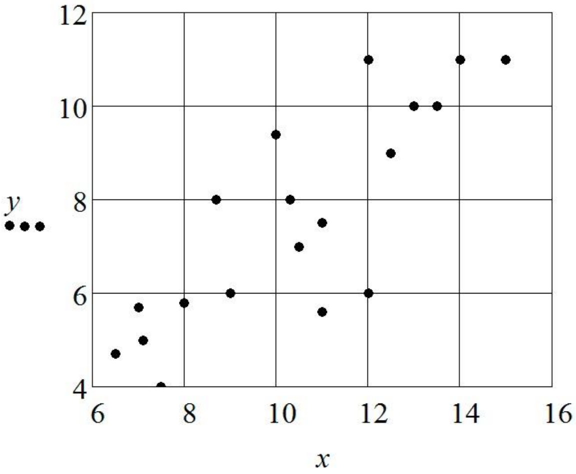
де X – незалежна змінна; Y – залежна змінна; β_0, β_1 – параметри моделі; ε – випадкова величина (збурення).

Випадкова величина ε включає вплив не врахованих у моделі факторів, випадкових похибок та особливостей вимірювання. Її присутність у моделі викликана трьома причинами: специфікацією моделі, вибірковим характером вихідних даних, особливостями вимірювання змінних [1].

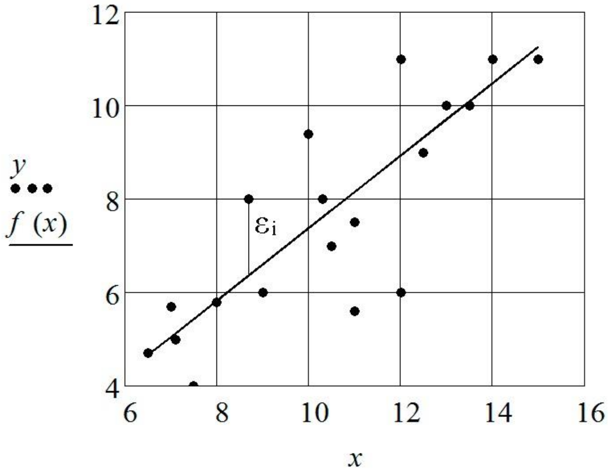
В основі оцінки рівняння парної лінійної регресії лежать вибірки значень двох змінних – незалежної X (2.1) та залежної Y (2.2). Геометричною інтерпретацією таких вибірок розміром n буде сукупність n точок на площині (рис. 3.1 *а*), яка називається *діаграмою розсіювання* чи *кореляційним полем*.

Задачею парного лінійного регресійного аналізу буде знаходження такої прямої, яка найкращим чином аналітично характеризує (апроксимує) кореляційне поле. Тобто потрібно мінімізувати деяку «міру» відхилень $g(\varepsilon_i)$ від лінії регресії (рис. 3.1 *б*).

Під *апроксимацією* тут необхідно розуміти аналітичний і (або) графічний опис деякого набору статистичних спостережень.



а



б

Рисунок 3.1 – Геометрична інтерпретація двох вибірок статистичних даних: *а* – кореляційне поле; *б* – кореляційне поле із лінією регресії

Як міру відхилення функції $y_i = \beta_0 + \beta_1 \cdot x_i$ від точок набору спостережень можна взяти [4]:

1) суму квадратів відхилень

$$F = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 \cdot x)]^2 . \quad (3.2)$$

2) суму модулів відхилень

$$F=\sum_{i=1}^{n}|\varepsilon_i|=\sum_{i=1}^{n}|y_i-(\beta_0+\beta_1\cdot x)|. \qquad (3.3)$$

3) функцію Хубера, яка при малих відхиленнях квадратична, а при великих – лінійна.

Із ряду причин, які будуть розглядатися нижче, найбільшого поширення як міра відхилення набула сума квадратів відхилень. Метод знаходження оцінок параметрів регресії за допомогою цієї функції дістав назву методу найменших квадратів.

3.2. Метод найменших квадратів

Згідно з методом найменших квадратів (МНК) оцінки невідомих параметрів β_0,β_1 обираються таким чином, щоб сума квадратів відхилень значень, знайдених за рівнянням $\hat{y}_i=\beta_0+\beta_1\cdot x_i$ від емпіричних значень y_i, була мінімальною [1–7]:

$$F=\sum_{i=1}^{n}\varepsilon_i^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}[y_i-(\beta_0+\beta_1\cdot x)]^2\to\min. \qquad (3.4)$$

Функціонал (3.4) також іноді називають *функцією незв'язності*.

На підставі необхідної умови екстремуму функції двох змінних β_0,β_1 прирівнюємо до нуля її частинні похідні:

$$\begin{cases}\dfrac{\partial F}{\partial\beta_0}=-2\cdot\sum_{i=1}^{n}[y_i-(\beta_0+\beta_1\cdot x_i)]=0;\\ \dfrac{\partial F}{\partial\beta_1}=-2\cdot\sum_{i=1}^{n}x_i\cdot[y_i-(\beta_0+\beta_1\cdot x_i)]=0.\end{cases}. \qquad (3.5)$$

Для подальших математичних викладок систему (1.15) доцільно записати у вигляді:

$$\begin{cases} \sum_{i=1}^{n}[y_i - (\beta_0 + \beta_1 \cdot x_i)] = 0; \\ \sum_{i=1}^{n} x_i \cdot [y_i - (\beta_0 + \beta_1 \cdot x_i)] = 0. \end{cases} \tag{3.6}$$

Розкриємо дужки та отримаємо стандартну форму нормальних рівнянь:

$$\begin{cases} \beta_0 \cdot n + \beta_1 \cdot \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i; \\ \beta_0 \cdot \sum_{i=1}^{n} x_i + \beta_1 \cdot \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i \cdot y_i. \end{cases} \tag{3.7}$$

Потім, розділивши обидві частини системи на n, отримаємо систему із двох алгебричних рівнянь із двома невідомими β_0, β_1:

$$\begin{cases} \beta_0 + \beta_1 \cdot \bar{x} = \bar{y}; \\ \beta_0 \cdot \bar{x} + \beta_1 \cdot \overline{x^2} = \overline{xy}; \end{cases} \tag{3.8}$$

Розв’язуючи систему (3.8) алгебричних рівнянь відносно β_0, β_1, одержимо:

$$\beta_0 = \bar{y} - \beta_1 \cdot \bar{x}; \tag{3.9}$$

$$\beta_1 = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{Cov(X, Y)}{Var(X)}. \tag{3.10}$$

Потрібно відзначити, що (3.9) та (3.10) не є параметрами парної лінійної регресії, а лише оцінками цих параметрів, отриманими за МНК. Реальні значення параметрів невідомі, оскільки на практиці працюють не із генеральними сукупностями, а із вибірками (2.1), (2.2) з них.

Обрання як міри відхилення ліній регресії від емпіричних точок обґрунтовується *теоремою Гаусса-Маркова* [1, 4]: при виконанні ряду вимог до вихідних вибірок статистичних спостережень, МНК-оцінки параметрів лінійної регресії є найбільш *ефективними*, тобто мають найменшу дисперсію в класі усіх лінійних незміщених оцінок.

Вимоги до вихідних вибірок статистичних спостережень будуть розглянуті в темі «Множинний лінійний кореляційно-регресійний аналіз», оскільки він є узагальненням парного лінійного кореляційно-регресійного аналізу на моделі із довільним числом незалежних змінних.

Величина $\hat{\beta}_j$ називається *незміщеною оцінкою* параметра β_j, якщо її вибіркове математичне очікування дорівнює параметру β_j генеральної сукупності $M[\hat{\beta}_j] = \beta_j$. Змістовно незміщеність оцінки означає, що вона не має систематичної похибки [4, 7].

Серед характеристик оцінок параметрів також потрібно відзначити *обґрунтованість* – властивість оцінки $\hat{\beta}_j$ параметра β_j наближатися до його реального значення при збільшенні обсягу n вибірки, за якою проводилася оцінка $\lim\limits_{n \to \infty} \hat{\beta}_j = \beta_j$ [4].

Оскільки МНК при виконанні ряду вимог до вихідних вибірок спостережень дає найбільш ефективні оцінки параметрів парної лінійної регресії, його алгоритм реалізований багатьма комп’ютерними програмами. Так у Microsoft Excel існує статистична функція «ЛИНЕЙН» та інструмент «Регресія» надбудови «Аналіз даних». У MathCAD – оператор *line(X,Y)*.

3.3. Економічний зміст параметрів парної лінійної регресії

Економічний зміст параметрів парної лінійної регресії визначається сутністю залежної X та незалежною Y змінними. Розглянемо декілька прикладів.

У підручнику [2] вивчається потреба підприємства в електроенергії Y залежно від обсягу продукції X, що випускається. Якщо потреба в електроенергії може бути оцінена рівнянням парної лінійної регресії

$\hat{Y} = \beta_0 + \beta_1 \cdot X$, то все споживання електроенергії $\hat{Y}$ можна розділити на дві частини – безпосередньо не пов'язане із виробництвом продукції β_0 та безпосередньо пов'язане з обсягом продукції, що випускається $(\beta_1 \cdot x)$, яке пропорційно зростає із збільшенням обсягу випуску X. Причому β_1 показує оцінку безпосередніх витрат електроенергії на кожну одиницю виробленої продукції.

За аналогією із цим прикладом можна запропонувати такий. Залежність витрат підприємства від обсягу виробництва взагалі характеризується нелінійною залежністю. Однак на певному невеликому відрізку вони можуть оцінюватися рівнянням регресії $\hat{Y} = \beta_0 + \beta_1 \cdot X$. Тоді витрати підприємства можуть бути розділені на умовно-постійні β_0, такі, що не змінюються із зміною обсягу виробництва (орендна плата, утримання адміністрації та ін.), і на умовно-змінні $(\beta_1 \cdot X)$, такі, що змінюються пропорційно зміні обсягу продукції (витрата матеріалу, заробітна плата працівників зайнятих безпосередньо на виробництві, та ін.). Причому тут β_1 – оцінка умовно-змінних витрат на одиницю виробленої продукції.

Наведені вище приклади ілюструють пряму залежність Y від X. Розглянемо зворотну залежність кількості бракованої продукції Y (%) від кількості робітників із спеціальною підготовкою X (%). Якщо така залежність буде описуватися рівнянням $\hat{Y} = \beta_0 - \beta_1 \cdot X$, то β_0 – оцінка кількості бракованої продукції при відсутності робітників із спеціальною підготовкою $x = 0$, β_1 – оцінка зменшення кількості (%) бракованої продукції при збільшенні кількості робітників із спеціальною підготовкою на 1 %.

Параметр β_0 не завжди буде мати економічний зміст, зокрема він не має економічного змісту, якщо він від'ємний. У [3] розглядається приклад, коли параметр β_0 не має економічного змісту, навіть в тому випадку, коли він позитивний.

Запитання для самоперевірки

1. Дати визначення регресійного аналізу.
2. Записати модель парної лінійної регресії та вказати її складові.

3. Розкрити поняття «кореляційне поле» та «апроксимація».

4. Навести приклади функцій нев'язності, що можуть бути застосовані для оцінки параметрів парної лінійної регресії.

5. Коротко описати ідею та основні положення МНК.

6. В чому полягає теорема Гауса-Маркова?

7. Економічний зміст параметрів парної лінійної регресії в загальному випадку.

8. Навести приклади економічного сенсу параметрів парної лінійної регресії в залежності від економічного сенсу залежної та незалежної змінних.

Список літератури

1. Кремер Н.Ш. Эконометрика : учебник для вузов / Н.Ш. Кремер, Б.А. Путко ; под ред. Н.Ш. Кремера. – М. : ЮНИТИ-ДАНА, 2007. – 311 с.
2. Елисеева И.И. Эконометрика : учеб. / И.И. Елисеева ; под ред. И.И. Елисеевой. – М. : Финансы и статистика, 2004. – 344 с.
3. Доугерти К. Введение в эконометрику / К. Доугерти ; пер. с англ. – М. : ИНФРА-М, 1999. – 402 с.
4. Магнус Я.Р. Эконометрика. Начальный курс : учеб. / Я.Р. Магнус, П.К. Катышев, А.А. Пересецкий. – 6-е изд., перераб. и доп. – М. : Дело, 2004. – 576 с.
5. Айвазян С.А. Прикладная статистика и основы эконометрики / С.А. Айвазян, В.С. Мхитарян. – М. : ЮНИТИ, 1998. – 432 с.
6. Джонстон Дж. Эконометрические методы / Дж. Джонстон ; пер. с англ. – М.: Статистика, 1980. – 444 с.
7. Лещинський О.Л. Економетрія : навч. посіб. для студ. вищ. навч. закл. / О.Л. Лещинський, В.В. Рязанцева, О.О. Юнькова. – К. : МАУП, 2000. – 208 с.
8. Кремер Н.Ш. Математическая статистика. – М.: Економическое образование, 1992. – 203 с.

ТЕМА 4. ОЦІНКА ЯКОСТІ МОДЕЛІ ПАРНОЇ ЛІНІЙНОЇ РЕГРЕСІЇ

Перевірка статистичної значущості параметрів регресії за критерієм Стю'дента. Побудова довірчих інтервалів оцінок параметрів регресії. Коефіцієнт детермінації. Перевірка статистичної значущості коефіцієнта детермінації за критерієм Фішера.

4.1. Перевірка статистичної значущості та довірчі інтервали оцінок параметрів регресії

Параметри регресії оцінюються на основі вибірки емпіричних величин, у значеннях яких суттєву роль відіграють випадкові фактори. Це призводить до наявності суттєвої випадкової складової в оцінці кожного із параметрів, тому вони потребують перевірки статистичної значущості та побудови довірчих інтервалів. З цією метою за кожним з параметрів визначається його стандартна помилка – m_{β_0}, m_{β_1} [1–7]. Стандартна помилка оцінки параметра регресії β_1 визначається за формулою:

$$m_{\beta_1} = \sqrt{\frac{S^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}, \tag{4.1}$$

де S^2 – залишкова дисперсія на один степінь свободи.

У разі нормального розподілу величин відхилень ε вона розраховується за формулою:

$$S^2 = \frac{\sum_{i=1}^{n}\varepsilon_i^2}{n-2} = \frac{\sum_{i=1}^{n}[y_i - (\beta_0 + \beta_1 \cdot x_i)]^2}{n-2}. \tag{4.2}$$

Далі визначається фактичне значення *t*-критерію Ст’юдента:

$$t_{\beta_1} = \frac{\beta_1}{m_{\beta_1}}, \tag{4.3}$$

яке потім порівнюється критичним (табличним), обраним при певному рівні вірогідності p, або значущості α, і числі степенів свободи $k = n - 2$.

Якщо фактичне значення *t*-критерію перевищує табличне:

$$t_{\beta_1} > t_{p,k}, \tag{4.4}$$

то гіпотезу про статистичну незначущість оцінки параметра β_1 регресії можна відхилити.

Довірчий інтервал оцінки параметра β_1 регресії визначається:

$$\beta_1 - t_{k,p} \cdot m_{\beta_1} \le b_1 \le \beta_1 + t_{k,p} \cdot m_{\beta_1}, \tag{4.5}$$

де b_1 – дійсне значення параметра β_1 регресії, яке могло б бути отримане при оцінці за генеральною сукупністю.

Стандартна помилка параметра β_0 визначається за формулою:

$$m_{\beta_0} = \sqrt{S^2 \cdot \frac{\sum_{i=1}^{n} x_i^2}{n \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2}}. \tag{4.6}$$

В іншому послідовність оцінювання статистичної значущості оцінки параметра β_0 і побудови її довірчого інтервалу не відрізняється від розглянутої вище для β_1.

Оскільки оцінка параметра регресії в економетричних дослідженнях має чітку економічну інтерпретацію, то довірчі межі інтервалу для неї не повинні містити суперечливих результатів, наприклад $-10 \le b_j \le 40$. Запис , що дійсне значення параметра регресії може приймати позитивні, від'ємні і навіть нульове значення, чого, скоріше за все, не може бути з економічної точки зору.

При перевірці статистичної значущості та побудові довірчих інтервалів параметрів регресії обов'язково необхідно вказувати вірогідність *p*, для якої визначалося $t_{p,k}$. Це пов'язано із тим, що для одного рівня вірогідності оцінка параметра регресії може бути статистично значущою, в той же час, як для вищого рівня вірогідності вона буде статистично незначущою. Межі довірчих інтервалів також будуть відрізнятися для різних значень вірогідності *p*.

4.2. Оцінка якості моделі парної лінійної регресії вцілому

Окрім перевірки статистичної значущості оцінок параметрів парної лінійної регресії, важливо також визначити якість моделі в цілому. Однією з найбільш ефективних оцінок адекватності регресійної моделі, мірою якості рівняння регресії, характеристикою її прогностичної сили є *коефіцієнт детермінації*. Цей коефіцієнт показує частину дисперсії, пояснену рівнянням регресії [3]:

$$R^2 = \frac{Var(\hat{Y})}{Var(Y)}. \qquad (4.7)$$

Іншою формою подання формули (4.7) може бути така:

$$R^2 = 1 - \frac{Var(\varepsilon)}{Var(Y)}. \qquad (4.8)$$

У разі парної лінійної регресійної моделі коефіцієнт детермінації рівний квадрату коефіцієнта кореляції:

$$R^2 = r^2. \qquad (4.9)$$

Чим ближче R^2 до одиниці, тим краще регресія апроксимує емпіричні дані, тим тісніше точки спостереження примикають до лінії регресії. Якщо $R^2 = 1$, то емпіричні точки, що характеризують незалежну X та залежну Y змінні, лежать на лінії регресії і між змінними існує лінійна функціональна залежність. Якщо $R^2 = 0$, то варіація залежної змінної повністю обумовлена дією неврахованих у моделі змінних, і лінія регресії паралельна осі абсцис. Зазначимо, що коефіцієнт R^2 є сенс розглядати тільки за наявності параметра β_0 в рівнянні регресії.

Якщо коефіцієнт детермінації відомий, то критерій його статистичної значущості (F-критерій Фішера) для парної лінійної регресії може бути записаний у вигляді:

$$F = \frac{R^2 \cdot (n-2)}{1-R^2}. \qquad (4.10)$$

Для заданого рівня значущості α, степенів свободи k_1 та k_2 знаходиться табличне (додаток В) значення критерію Фішера (у деяких джерелах він називається критерієм Фішера-Снедекора). Причому $k_1 = m-1$, де m – кількість змінних у моделі (для парної регресії $k_1 = 1$), $k_2 = n-2$. На сучасному етапі розвитку економетрії та інших наук, пов'язаних із математичною статистикою, критичне значення критерію Фішера можна визначити за допомогою статистичних функцій F.ОБР та F.ОБР.ПХ Microsoft Excel 2010. При використанні цієї функції F.ОБР.ПХ у рядку «Вірогідність» діалогового вікна потрібно зазначати не вірогідність p, а рівень значущості α. Така невідповідність довідки за функцією, скоріше за все, пов'язана із неточністю перекладу інтерфейсу статистичних функції із англійської мови на російську. В більш старих версіях Microsoft Excel використовується функція FРАСПОБР, у рядку «Вірогідність» діалогового вікна якої необхідно вказувати не вірогідність p, а рівень значущості α.

У MathCAD критичне значення F-критерію Фішера можна отримати за допомогою оператора $qF(p, k_1, k_2)$.

Якщо $F > F_{кр.}$, то для вірогідності p коефіцієнт детермінації статистично значущий.

Після побудови та дослідження якості моделі парної лінійної регресії можна перейти до прогнозування розвитку економічного явища чи процесу, який описує ця модель.

Запитання для самоперевірки

1. Що таке залишкова дисперсія та стандартні помилки параметрів регресії?
2. Яким чином здійснюється перевірка статистичної значущості параметрів регресії за критерієм Стю'дента?
3. Як будується довірчі інтервали оцінок параметрів регресії?
4. Дати визначення коефіцієнту детермінації.
5. Алгоритм перевірка статистичної значущості коефіцієнта детермінації за критерієм Фішера.
6. Який висновок можна зробити за значенням коефіцієнта детермінації.
7. Назвати комп’ютерні засоби перевірки статистичної значущості параметрів регресії та коефіцієнта детермінації.

Список літератури

1. Кремер Н.Ш. Эконометрика : учебник для вузов / Н.Ш. Кремер, Б.А. Путко ; под ред. Н.Ш. Кремера. – М. : ЮНИТИ-ДАНА, 2007. – 311 с.
2. Елисеева И.И. Эконометрика : учеб. / И.И. Елисеева ; под ред. И.И. Елисеевой. – М. : Финансы и статистика, 2004. – 344 с.
3. Доугерти К. Введение в эконометрику / К. Доугерти ; пер. с англ. – М. : ИНФРА-М, 1999. – 402 с.
4. Магнус Я.Р. Эконометрика. Начальный курс : учеб. / Я.Р. Магнус, П.К. Катышев, А.А. Пересецкий. – 6-е изд., перераб. и доп. – М. : Дело, 2004. – 576 с.
5. Айвазян С.А. Прикладная статистика и основы эконометрики / С.А. Айвазян, В.С. Мхитарян. – М. : ЮНИТИ, 1998. – 432 с.
6. Джонстон Дж. Эконометрические методы / Дж. Джонстон ; пер. с англ. – М.: Статистика, 1980. – 444 с.
7. Лещинський О.Л. Економетрія : навч. посіб. для студ. вищ. навч. закл. / О.Л. Лещинський, В.В. Рязанцева, О.О. Юнькова. – К. : МАУП, 2000. – 208 с.
8. Кремер Н.Ш. Математическая статистика. – М.: Економическое образование, 1992. – 203 с.

ТЕМА 5. ПРОГНОЗУВАННЯ НА ОСНОВІ МОДЕЛІ ПАРНОЇ ЛІНІЙНОЇ РЕГРЕСІЇ

Точковий та інтервальний прогноз. Екстраполяція. Довірчий інтервал для умовного математичного очікування. Довірчий інтервал для прогнозів індивідуальних значень.

У прогнозних розрахунках за рівнянням регресії визначається прогнозне значення залежної змінної, при деякому прогнозному значенні незалежної змінної $x_{\text{пр}}$. Прогноз поділяють на *точковий* (коли оцінкою прогнозної величини є число) та *інтервальний* (коли оцінкою є інтервал).

Точковою оцінкою залежної змінної вважають її умовне математичне очікування $M_{x_{\text{пр}}}(y)$, яке ототожнюють [1] із груповою середньою $\hat{y}_{x_{\text{пр}}}$, яку знаходять за рівнянням регресії:

$$\hat{y}_{x_{\text{пр}}} = \beta_0 + \beta_1 \cdot x_{\text{пр}} \,. \tag{5.1}$$

Дисперсію групової середньої оцінюють за формулою:

$$s^2_{\hat{y}_{x_{\text{пр}}}} = s \cdot \left(\frac{1}{n} + \frac{\left(x_{\text{пр}} - \bar{x}\right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\right). \tag{5.2}$$

Тоді стандартна помилка групової середньої:

$$s_{\hat{y}_{x_{\text{пр}}}} = \sqrt{s^2_{\hat{y}_{x_{\text{пр}}}}}. \tag{5.3}$$

Тоді p %-й довірчий інтервал для умовного математичного очікування становить:

$$\hat{y}_{x_{\text{пр}}} - s_{\hat{y}_{x_{\text{пр}}}} \cdot t_{p;k} \le M_{x_{\text{пр}}}(y) \le \hat{y}_{x_{\text{пр}}} + s_{\hat{y}_{x_{\text{пр}}}} \cdot t_{p;k}. \tag{5.4}$$

Побудована довірча область для $M_{x_{\text{пр}}}(y)$ умовного математичного очікування, але не окремих можливих прогнозних значень залежної змінної, які відхиляються від середньої. Тому при визначенні довірчого інтервалу для індивідуальних значень $y^*_{x_{\text{пр}}}$ залежної змінної необхідно враховувати іще одне джерело дисперсії – розсіювання навколо лінії регресії. Це реалізується шляхом включення в оцінку дисперсії $s^2_{\hat{y}_{x_{\text{пр}}}}$ величини s^2. У результаті оцінка дисперсії індивідуальних значень $y^*_{x_{\text{пр}}}$ дорівнює:

$$s^2_{y^*_{x_{\text{пр}}}} = s^2 \cdot \left(1 + \frac{1}{n} + \frac{\left(x_{\text{пр}} - \bar{x}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right). \tag{5.5}$$

Тоді стандартна помилка індивідуальних значень становить:

$$s_{y^*_{x_{\text{пр}}}} = \sqrt{s^2_{y^*_{x_{\text{пр}}}}}. \tag{5.6}$$

Для прогнозів індивідуальних значень p %-й довірчий інтервал $y^*_{x_{\text{пр}}}$ буде становити:

$$\hat{y}_{x_{\text{пр}}} - s_{y^*_{x_{\text{пр}}}} \cdot t_{p;k} \le y^*_{x_{\text{пр}}} \le \hat{y}_{x_{\text{пр}}} + s_{y^*_{x_{\text{пр}}}} \cdot t_{p;k}. \tag{5.7}$$

Чим далі $x_{\text{пр}}$ відхиляється від середнього за вибіркою значення незалежної змінної $\bar{x}$, тим менш точним буде прогноз і ширшими довірчі інтервали для умовного математичного очікування $M_{x_{\text{пр}}}(y)$ та індивідуальних значень $y^*_{x_{\text{пр}}}$. Тому *екстраполяція* регресії, тобто її використання поза межами дослідженого діапазону значень незалежної змінної, може призвести до значної похибки. Навіть у тому випадку, якщо така екстраполяція виправдана для розглянутої змінної, виходячи із економічного змісту задачі, що розв'язується.

При прогнозуванні розвитку економічних явищ чи процесів на основі економетричних моделей взагалі та парних лінійних регресійних моделей зокрема потрібно пам'ятати про складність та імовірнісний характер цих явищ та процесів. Тому отримані прогнози рекомендується використовувати лише як важливий допоміжний матеріал при прийнятті управлінського рішення, пов'язаного із прогнозом розвитку економічного явища чи процесу.

На економічні та соціально-економічні явища та процеси сукупно та одночасно діє значна кількість факторів різної природи, а модель парної регресії враховує лише один найбільш важливий, на думку дослідника, фактор. Більш ефективним інструментом опису вказаних вище явищ та процесів можуть служити множинні регресійні моделі, що одночасно враховують декілька факторів.

Запитання для самоперевірки

1. Назвати відмінності між точковим та інтервальним прогнозом.
2. Розкрити поняття екстраполяції рівняння регресії.
3. Побудова довірчого інтервалу для умовного математичного очікування.
4. Побудова довірчого інтервалу для прогнозів індивідуальних значень.
5. Відмінність між довірчими інтервалами для умовного математичного очікування та для прогнозів індивідуальних значень.

Список літератури

6. Кремер Н.Ш. Эконометрика : учебник для вузов / Н.Ш. Кремер, Б.А. Путко ; под ред. Н.Ш. Кремера. – М. : ЮНИТИ-ДАНА, 2007. – 311 с.
7. Елисеева И.И. Эконометрика : учеб. / И.И. Елисеева ; под ред. И.И. Елисеевой. – М. : Финансы и статистика, 2004. – 344 с.
8. Доугерти К. Введение в эконометрику / К. Доугерти ; пер. с англ. – М. : ИНФРА-М, 1999. – 402 с.
9. Магнус Я.Р. Эконометрика. Начальный курс : учеб. / Я.Р. Магнус, П.К. Катышев, А.А. Пересецкий. – 6-е изд., перераб. и доп. – М. : Дело, 2004. – 576 с.
10. Айвазян С.А. Прикладная статистика и основы эконометрики / С.А. Айвазян, В.С. Мхитарян. – М. : ЮНИТИ, 1998. – 432 с.
11. Джонстон Дж. Эконометрические методы / Дж. Джонстон ; пер. с англ. – М.: Статистика, 1980. – 444 с.
12. Лещинський О.Л. Економетрія : навч. посіб. для студ. вищ. навч. закл. / О.Л. Лещинський, В.В. Рязанцева, О.О. Юнькова. – К. : МАУП, 2000. – 208 с.
13. Кремер Н.Ш. Математическая статистика. – М.: Економическое образование, 1992. – 203 с.

ТЕМА 6. МНОЖИННИЙ КОРЕЛЯЦІЙНИЙ АНАЛІЗ

Оцінка лінійних взаємозв'язків між економічними змінними за парними та частими коефіцієнтами кореляції. Множинний коефіцієнт кореляції.

6.1. Оцінка лінійних взаємозв'язків між економічними змінними за парними коефіцієнтами кореляції

Економічні величини, як правило визначаються декількома одночасно і сукупно діючими факторами, причому ці фактори можуть бути взаємопов'язані. Для досліджень таких процесів можна використати методи множинного кореляційно-регресійного аналізу, який є узагальненням парного кореляційно-регресійного аналізу для довільного числа змінних.

Уявімо собі ситуацію, коли необхідно визначити тісноту взаємозв'язків між кількістю змінних $(m+1)$, представлених вибірками спостережень:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_i \\ \dots \\ y_n \end{pmatrix}, \tag{6.1}$$

$$x_1 = \begin{pmatrix} x_{11} \\ x_{21} \\ \dots \\ x_{i1} \\ \dots \\ x_{n1} \end{pmatrix}, \tag{6.2}$$

$$x_m = \begin{pmatrix} x_{1m} \\ x_{2m} \\ \dots \\ x_{im} \\ \dots \\ x_{nm} \end{pmatrix}, \tag{6.3}$$

де m – кількість незалежних змінних моделі.

Припустимо, що економічна суть вибірок (6.1), (6.2), (6.3) дозволяє нам визначити, що y є випадковою величиною, яка залежить від величин $x_1,...,x_m$. Оцінити тісноту лінійних взаємозв'язків між змінними дозволять нам парні коефіцієнти кореляції, які зручно представити у вигляді відповідної матриці:

$$q = \begin{pmatrix} 1 & r_{yx_1} & \dots & r_{yx_m} \\ r_{x_1y} & 1 & \dots & r_{x_1x_m} \\ \dots & \dots & \dots & \dots \\ r_{x_my} & r_{x_mx_1} & \dots & 1 \end{pmatrix}. \tag{6.4}$$

Матриця (6.4) є симетричною, причому по діагоналі матриці знаходяться одиниці. Парні коефіцієнти кореляції розраховуються тут за формулою (2.3). На сучасному етапі розвитку економетрії та інших наук, пов'язаних із математичною статистикою, найбільш доступним та зручним засобом побудови матриці (6.4) є інструмент «Кореляція» надбудови «Аналіз даних» Microsoft Excel.

За матрицею парних коефіцієнтів кореляції можна оцінити тісноту зв'язку залежної змінної y із кожною із незалежних змінних $x_1,...,x_m$. Також за (6.4) можна виявити *мільтиколінеарність* – тісний кореляційний зв'язок $(r > 0{,}7)$ між незалежними змінними $x_1,...,x_m$ на основі яких планується будувати модель. При виявленні мультиколінеарності будувати модель множинної лінійної регресії із використанням корельованих незалежних змінних за МНК не рекомендується, оскільки порушуються передумови теореми Гауса-Маркова.

Порядок перевірки статистичної значущості кожного із парних коефіцієнтів кореляції матриці (6.4) наведений в темі 2.

6.2. Оцінка лінійних взаємозв'язків між економічними змінними за часними коефіцієнтами кореляції

Парний коефіцієнт кореляції r_{jk} між двома вибірками економічних величин j та k враховує вплив інших змінних моделі, на тісноту звязку між j та k. У зв'язку із цим часто виникає необхідність визначити часну кореляцію між змінними при виключенні впливу інших змінних, на основі яких планується побудова моделі.

Часні коефіцієнти кореляції характеризують тісноту взаємозв'язку між двома змінними j та k множинній лінійної регресії без впливу інших змінних і в загальному випадку можуть бути обчислені як:

$$r_{чjk} = \frac{-q_{jk}}{\sqrt{q_{jj} \cdot q_{kk}}}, \tag{6.5}$$

де q_{jj}, q_{kk} – алгебраїчні доповнення елементів r_{jj}, r_{kk} матриці парних коефіцієнтів кореляції.

Для моделі із двома незалежними змінними x_1та x_2 часні коефіцієнти кореляції обчислюється за формулами:

$$r_{чyx_1} = \frac{r_{yx_1} - r_{yx_2} \cdot r_{x_1x_2}}{\sqrt{(1 - r_{yx_2}^2) \cdot (1 - r_{x_1x_2}^2)}}, \tag{6.6}$$

$$r_{чyx_2} = \frac{r_{yx_2} - r_{yx_1} \cdot r_{x_1x_2}}{\sqrt{(1 - r_{yx_1}^2) \cdot (1 - r_{x_1x_2}^2)}}, \tag{6.7}$$

$$r_{чx_1x_2} = \frac{r_{x_1x_2} - r_{yx_1} \cdot r_{yx_2}}{\sqrt{(1 - r_{yx_1}^2) \cdot (1 - r_{yx_2}^2)}}, \tag{6.8}$$

де $r_{yx_2}, r_{yx_1}, r_{x_1x_2}$ – парні коефіцієнти кореляції.

Як і парні коефіцієнти кореляції, часні можуть приймати значення від – 1 до 1.

Часний коефіцієнт кореляції вимагає перевірки статистичної значущості за критерієм Ст'юдента, фактичне значення якого розраховується за формулою:

$$t_{чjk} = \frac{r_{чjk} \cdot \sqrt{n - (m + 1)}}{\sqrt{1 - r_{чjk}^2}}. \tag{6.9}$$

Критичне значення критерію Ст'юдента отримується для певної вірогідності p (або рівня значущості α) та ступенів свободи:

$$f = n - (m + 1). \tag{6.10}$$

Якщо $\left|t_{чjk}\right| > t_{p,f}$ частний коефіцієнт кореляції статистично значуще відрізняється від 0 при вірогідності p. В противному випадку часний коефіцієнт кореляції статистично не значущий при цій вірогідності.

При реалізації множинного кореляційно-регресійного аналізу доцільно будувати матрицю часних коефіцієнтів кореляції:

$$q_ч = \begin{pmatrix} 1 & r_{чyx_1} & \dots & r_{чyx_m} \\ r_{чx_1y} & 1 & \dots & r_{чx_1x_m} \\ \dots & \dots & \dots & \dots \\ r_{чx_my} & r_{чx_mx_1} & \dots & 1 \end{pmatrix}. \tag{6.11}$$

6.3. Множинний коефіцієнт кореляції

Множинний коефіцієнт кореляції характеризує тісноту статистичного взаємозв'язку між залежної змінною і всіма незалежними змінними, які використовуються при побудові моделі. Одною із найбільш зручних формул для обчислення множинного коефіцієнта кореляції, є наступна:

$$r_{мн} = \sqrt{1 - \frac{\Delta q}{\Delta q_{xx}}}, \tag{6.12}$$

де $\Delta q, \Delta q_{xx}$ – визначники матриць q, q_{xx}, відповідно;

q_{xx} – матриця складена із парних коефіцієнтів кореляції між незалежними змінними на основі яких планується будувати модель.

Зручність формули (6.12) полягає в простоті обчислення визначника матриці за допомогою функції «МОПРЕД» в MS Excel або оператора $|x|$ в MathCAD. Інструмент «Регресія» надбудови «Аналіз даних» Microsoft Excel розраховує множинний коефіцієнт кореляції в числі іншої інформації, яка характеризує якість побудованої моделі.

Коефіцієнт множинної кореляції змінюється в межах від 0 до +1 і тому не призначений для визначення напрямку зв'язку між залежною та незалежними змінними. Чим ближче множинний коефіцієнт кореляції до одиниці, тим сильніше взаємозв'язок між залежною і всім набором незалежних змінних, і, навпаки, чим ближче множинний коефіцієнт кореляції до нуля, тим слабкіше взаємозв'язок між досліджуваними змінними.

Перевірка гіпотези про незначущість множинного коефіцієнта кореляції здійснюється за допомогою *F*-критерію Фішера через коефіцієнт детермінації (див. тему __).

Розглянуті вище коефіцієнти кореляції дозволяють всебічно оцінити тісноту зв'язків між змінними на основі яких планується будувати модель множинної лінійної регресії.

Запитання для самоперевірки

1. Оцінка тісноти лінійного зв'язку між декількома змінними за матрицею парних коефіцієнтів кореляції.
2. Мультиколінеарність.
3. Матриця часних коефіцієнтів кореляції. Її подібності та відмінності відносно матриці парних коефіцієнтів кореляції.
4. Множинний коефіцієнт кореляції.

Список літератури

1. Кремер Н.Ш., Путко Б.А. Эконометрика: Учебник для вузов / Под. ред. проф. Н.Ш. Кремера. – М.: ЮНИТИ-ДАНА, 2007. – 311 с.
2. Наконечний С.І., Терещенко Т.О., Романюк Т.П. Економетрія: Підручник. – К.: КНЕУ, 2005. – 520 с.
3. Эконометрика: Учебник / Под ред. И.И. Елисеевой. – М.: Финансы и статистика, 2004. – 344 с.
4. Магнус Я.Р., Катышев П.К., Пересецкий А.А. Эконометрика. Начальный курс: Учеб. – 6-е изд., перераб. и доп. – М.: Дело, 2004. – 576 с.
5. Введение в эконометрику / К. Доугерти ; пер. с англ. – М: ИНФРА-М, 1999. – 402 с.
6. Джонстон Дж. Эконометрические методы / Дж. Джонстон ; пер. с англ. – М.: Статистика, 1980. – 444 с.

ТЕМА 7. КЛАСИЧНА НОРМАЛЬНА ЛІНІЙНА МОДЕЛЬ МНОЖИННОЇ РЕГРЕСІЇ

Загальний вигляд моделі множинної лінійної регресії. Передумови застосування МНК для оцінки вектору параметрів моделі множинної лінійної регресії. Гомо- та гетероскедастичність. Мультиколінеарність. Автокореляція. Використання МНК для оцінки параметрів множинної лінійної регресії. Стандартизовані параметри регресії та коефіцієнти еластичності.

7.1. Загальний вигляд моделі множинної лінійної регресії

Лінійна модель множинної регресії є узагальненням моделі парної лінійної регресії для випадку довільної кількості незалежних змінних моделі:

$$\hat{y}_i = \beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + \ldots + \beta_m \cdot x_{im} + \varepsilon_i . \qquad (7.1)$$

Як правило, при дослідженні економічних явищ та процесів за допомогою таких моделей, кількість m керованих змінних в них не перевищує 6-7, причому найбільш розповсюдженими є моделі множинної лінійної регресії із 2-3 незалежними змінними.

Як і у випадку парної лінійної регресії, МНК є найбільш розповсюдженим методом оцінки вектору параметрів множинної лінійної регресії:

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \ldots \\ \beta_m \end{pmatrix}; \qquad (7.2)$$

7.2. Передумови застосування МНК для оцінки вектору параметрів моделі множинної лінійної регресії

Модель (7.1) буде називатися *класичною нормальною лінійною моделлю множинної регресії*, якщо залежна змінна y, відхилення ε, незалежні змінні $x_1,\ldots,x_m$ задовольняють наступним вимогам:

1. У моделі відхилення ε (або залежна змінна y) є величина випадкова, а незалежні змінні $x_1,...,x_m$, – величини невипадкові. При цьому передбачається, що серед вибіркових значень $x_1...,x_m$ не всі однакові.

2. Математичне очікування відхилення ε_i, дорівнює нулю:

$$M(\varepsilon_i) = 0. \tag{7.3}$$

Іншими словами математичне очікування залежної змінної y_i визначається рівнянням регресії.

3. Дисперсія відхилень ε_i (або залежної змінної y_i) має бути постійною для будь-якої групи спостережень:

$$D(\varepsilon_i) = const; \tag{7.4}$$

$$D(y_i) = const; \tag{7.5}$$

для будь-якого $i = 1...n$.

Рівняння (7.4), (7.5) рівнозмінності дисперсії відхилень залежної змінної є умовою *гомоскедастичності*. Натомість явище при котрому дисперсія відхилень залежної змінної відрізняється для різних груп спостережень, тобто умови (7.4), (7.5) не виконуються, називається *гетероскедастичністю*.

4. Послідовні частини $y_1,...,y_i$ та $y_{i+1},...,y_n$ однієї вибірки $y_1,...,y_i,...,y_n$ незалежної змінної не корелюються між собою. Відповідно відхилення $\varepsilon_1,...,\varepsilon_i$ й $\varepsilon_{i+1},...,\varepsilon_n$ теж не корельовано між собою, тобто *відсутня автокореляція*. *Автокореляцією* називається явище корельованості послідовних частин однієї вибірки спостережень.

5. Відсутня *мільтиколінеарність* – тісний кореляційний зв'язок $(r > 0{,}7)$ між незалежними змінними $x_1,...,x_m$.

6. Збурювання ε_i, (або залежна змінна y_i) є нормально розподілена випадкова величина.

При виконанні передумов 1-5 справедлива *теорема Гауса-Маркова*: оцінка вектору β (7.2) параметрів множинної лінійної регресії (7.1), отримана за допомогою МНК є найбільш ефективною, тобто має найменшу дисперсію в класі лінійних незміщених оцінок.

Для використання МНК передумова 6 не є обов'язковою, вимога нормального розподілення ε_i (y_i) потрібна для оцінки точності параметрів та рівняння регресії вцілому. Якщо передумова 6 не виконується, модель (7.1) може називатися просто *класичною лінійною моделлю множинної регресії*. Доказ теореми Гауса-Маркова представлений у багатьох підручниках із економетрії, наприклад [1].

Модель парної лінійної регресії можна розглядати, як часний випадок множинної лінійної регресії при наявності лише одної незалежної змінної. Таким чином, для парної лінійної регресії також справедлива теорема Гауса-Маркова, якщо виконуються передумови 1-4. Причому при їх виконанні модель буде називатися класичною лінійною моделлю парної регресії, а при додатковому виконанні передумови 6 – класичною нормальною лінійною моделлю парної регресії. Очевидно, що передумова 5, відсутності мультиколінеарності, для парної лінійної регресії не актуальна, оскільки тут присутня лише одна незалежна змінна.

7.3. Використання МНК для оцінки параметрів множинної лінійної регресії

Розглянемо узагальнення МНК для випадку множинної лінійної регресії:

$$F = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \to \min, \tag{7.6}$$

$$F = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + \ldots + \beta_m \cdot x_{im})]^2 \to \min. \tag{7.7}$$

Включення в регресійну модель нових незалежних змінних ускладнює обчислення. Це приводить до доцільності використання матричних позначень.

Окрім вектору спостережень залежної змінної (6.1) та вектору параметрів регресії (7.2), запишемо вектор оцінок відхилень:

$$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{pmatrix}. \tag{7.8}$$

Якщо планується побудова моделі у вигляді (7.1), то вектори спостережень незалежних змінних (6.2) та (6.3) об'єднаємо у матрицю виду:

$$X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1m} \\ 1 & x_{21} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots \\ 1 & x_{n1} & \dots & x_{nm} \end{pmatrix}; \tag{7.9}$$

Наявність в (7.9) першого стовпчику, що складається із одиниць пояснюється тим, що в моделі (7.1) параметр β_0 помножений на одиницю, в той час як інші параметри $\beta_1,\dots,\beta_m$ помножені на відповідні незалежні змінні $x_1,\dots,x_m$.

Якщо приймають $\beta_0 = 0$ тобто будують модель множинної лінійної регресії виду:

$$\hat{y}_i = \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + \dots + \beta_m \cdot x_{im} + \varepsilon_i, \tag{7.10}$$

то матрицю (7.9) використовують у вигляді:

$$X = \begin{pmatrix} x_{11} \dots x_{1m} \\ x_{21} \dots x_{2m} \\ \dots \dots \dots \\ x_{n1} \dots x_{nm} \end{pmatrix}. \tag{7.11}$$

Враховуючи матричні позначення (6.1), (7.2), (7.8), (7.9) або (7.11) перепишемо модель множинної лінійної регресії (7.1) або (7.10) в матричній формі:

$$Y = X \cdot \beta + \varepsilon . \tag{7.12}$$

Тоді функція незв'язності (7.7) матиме вигляд:

$$F = \sum_{i=1}^{n} \varepsilon_i^2 = \left[\varepsilon^T \cdot \varepsilon\right] = \left[(Y - X \cdot \beta)^T \cdot (Y - X \cdot \beta)\right] \to \min , \tag{7.13}$$

В результаті відповідних математичних перетворень [1] отримаємо рівняння для оцінки вектору параметрів множинної лінійної регресії:

$$\beta = (X^T \cdot X)^{-1} \cdot (X^T \cdot Y) . \tag{7.14}$$

Одним із сучасних практичних способів оцінки вектору параметрів β за МНК може бути запис вектору спостережень залежної змінної (6.1) та матриці спостережень незалежної змінної (7.9) або (7.11) в MathCAD із подальшим використанням формули (7.14). Іншими способами отримати оцінки β за МНК можуть бути статистична функція ЛИНЕЙН або інструмент «Регресія» надбудови «Аналіз даних» Microsoft Excel, причому ці засоби дозволяють не тільки оцінити параметри моделі множинної лінійної регресії, а і отримати додаткову інформацію про якість отриманої моделі.

7.4 Стандартизовані параметри регресії та коефіцієнти еластичності

На практиці часто буває необхідно порівняння впливу на залежну змінну різних незалежних змінних, коли останні виражаються різними одиницями вимірювання. У цьому випадку використовують стандартизовані параметри регресії β'_j і коефіцієнти еластичності E_j.

Стандартизований параметр регресії β'_j показує на скільки величин s_y зміниться в середньому залежна змінна y при збільшенні тільки *j-й* незалежної змінної на s_{x_j}:

$$\beta'_j = \beta_j \cdot \frac{s_{x_j}}{s_y}; \qquad (7.15)$$

де s_{x_j} – середньоквадратичне відхилення змінної x_j;

s_y – середньоквадратичне відхилення змінної y;

j – номер змінної в рівнянні регресії, $j = 1...m$.

Середньоквадратичне відхилення змінної x_j розраховується за формулою:

$$s_{x_j} = \sqrt{\frac{\sum_{i=1}^{n} x_{ij}^2}{n} - (\bar{x}_j)^2}; \qquad (7.16)$$

Середньоквадратичне відхилення змінної y розраховується за формулою:

$$s_y = \sqrt{\frac{\sum_{i=1}^{n} y_i^2}{n} - (\bar{y})^2}. \qquad (7.17)$$

Інший підхід до обчислення оцінок стандартизованих параметрів регресії полягає в переході до *стандартизованих змінних* за формулами:

$$t_{x_{ij}} = \frac{x_{ij} - \bar{x}_j}{s_{x_j}}; \qquad (7.18)$$

$$t_{y_i} = \frac{y_i - \bar{y}}{s_y}. \qquad (7.19)$$

Середнє значення стандартизованих змінних дорівнює нулю, а середньоквадратичне відхилення одиниці.

Після перетворень по формулах (7.18) і (7.19) рівняння множинної регресії може оцінюватися за МНК у вигляді:

$$t_{y_i} = \beta_0^{'} + \beta_1^{'} \cdot t_{x_{i1}} + \beta_2^{'} \cdot t_{x_{i2}} + \ldots + \beta_m^{'} \cdot t_{x_{im}}. \quad (7.20)$$

Отримані МНК-оцінки стандартизованих параметрів регресії $\beta_0^{'}, \beta_1^{'}, \beta_2^{'}, \ldots, \beta_m^{'}$ можна перевірити за формулою (7.15), попередньо розрахувавши за МНК звичайні параметри регресії $\beta_0, \beta_1, \beta_2, \ldots, \beta_m$.

Всі стандартизовані параметри регресії мають однакову розмірність і можуть бути співставленні між собою. Їх можна використовувати для відбору незалежних змінних при специфікації моделі лінійної регресії. При побудові моделі не використовуються незалежні змінні із найменшим по модулі значенням $\left|\beta_j^{'}\right|$.

Коефіцієнт еластичності E_j показує на скільки відсотків (від середньої) зміниться в середньому y при збільшенні x_j тільки на 1%:

$$E_j = \beta_j \cdot \frac{\bar{x}_j}{\bar{y}}. \quad (7.21)$$

Таким чином, коефіцієнт еластичності також можна використовувати для відбору незалежних змінних при специфікації моделі лінійної регресії. Чим більше коефіцієнт еластичності, тим більше вплив відповідної йому незалежної змінної на залежну.

Запитання для самоперевірки

1. Записати модель множинної лінійної регресії в загальному вигляді та вказати її складові.
2. Як співвідносяться моделі парної та множинної лінійної регресії?

3. Коротко охарактеризувати вимоги до вихідних статистичних даних, які мають виконуватись для коректного застосування МНК.

4. Розкрити поняття гомо- та гетероскдастичності.

5. Розкрити поняття автокореляції.

6. Розкрити поняття мультиколінеарності.

7. Класична нормальна лінійна модель множинної регресії.

8. Теорема Гауса-Маркова для множинної лінійної регресії.

9. Особливості застосування МНК для оцінки параметрів класичної лінійної моделі множинної регресії.

10. Комп'ютерні засоби для оцінки вектору параметрів моделі множинної лінійної регресії за МНК.

11. Назвати мету отримання стандартизованих параметрів множинної регресії та коефіцієнтів еластичності.

12. Назвати способи оцінки стандартизованих параметрів множинної лінійної регресії.

13. Коефіцієнти еластичності для незалежних змінних моделі множинної лінійної регресії.

Список літератури

1. Кремер Н.Ш., Путко Б.А. Эконометрика: Учебник для вузов / Под. ред. проф. Н.Ш. Кремера. – М.: ЮНИТИ-ДАНА, 2007. – 311 с.

2. Наконечний С.І., Терещенко Т.О., Романюк Т.П. Економетрія: Підручник. – К.: КНЕУ, 2005. – 520 с.

3. Эконометрика: Учебник / Под ред. И.И. Елисеевой. – М.: Финансы и статистика, 2004. – 344 с.

4. Магнус Я.Р., Катышев П.К., Пересецкий А.А. Эконометрика. Начальный курс: Учеб. – 6-е изд., перераб. и доп. – М.: Дело, 2004. – 576 с.

5. Введение в эконометрику / К. Доугерти ; пер. с англ. – М: ИНФРА-М, 1999. – 402 с.

6. Джонстон Дж. Эконометрические методы / Дж. Джонстон ; пер. с англ. – М.: Статистика, 1980. – 444 с.

ТЕМА 8. ОЦІНКА ЯКОСТІ МОДЕЛІ МНОЖИННОЇ ЛІНІЙНОЇ РЕГРЕСІЇ ТА ПРОГНОЗУВАННЯ НА ЙОГО ОСНОВІ

Статистична значущість і довірчі інтервали оцінок параметрів регресії. Оцінка значущості моделі множинної регресії вцілому. Прогнозування на основі моделей множинної лінійної регресії.

8.1. Статистична значущість і довірчі інтервали оцінок параметрів регресії

МНК-оцінки параметрів множинної лінійної регресії (7.2) є випадковою величиною, тому потребують перевірки статистичної значущої відмінності від 0 за критерієм Ст'юдента. Фактичне значення цього критерію для параметру β_j визначається за формулою:

$$t_{\beta_j} = \frac{\beta_j}{s_{\beta_j}}. \tag{8.1}$$

де s_{β_j} – стандартна помилка параметру β_j.

Стандартна помилка параметру β_j розраховується за формулою:

$$s_{\beta_j} = s \cdot \sqrt{[(X^T \cdot X)^{-1}]_{jj}}, \tag{8.2}$$

де s – стандартна помилка регресії;

$[(X^T \cdot X)^{-1}]_{jj}$ – діагональний елемент матриці $(X^T \cdot X)^{-1}$.

Стандартна помилка множинної лінійної регресії обчислюється за формулою:

$$s = \sqrt{\frac{\sum_{i=1}^{n} \varepsilon_i^2}{n-(m+1)}}, \tag{8.3}$$

або

$$s = \sqrt{\frac{\varepsilon^T \cdot \varepsilon}{n-(m+1)}}. \tag{8.4}$$

Величина $s_{\beta_j}^2$ називається оцінкою дисперсії параметру регресії, а $\sigma_\varepsilon^2 = s^2$ – оцінкою дисперсії помилки регресії.

Фактичне значення критерію Ст'юдента порівнюється із табличним, отриманим для певної вірогідності p (або рівня значущості α) та ступенів свободи f. Якщо $\left|t_{\beta_j}\right| > t_{p,f}$ параметр регресії статистично значуще відрізняється від 0 при вірогідності p. В противному випадку параметр регресії статистично не значущий при цій вірогідності.

Довірчий інтервал для параметра β_j будується у вигляді:

$$(\beta_j - s_{\beta_j} \cdot t_{p,f}) \le \beta_j \le (\beta_j + s_{\beta_j} \cdot t_{p,f}); \tag{8.5}$$

або

$$\beta_j \pm s_{\beta_j} \cdot t_{p,f}. \tag{8.6}$$

Зручним практичним засобом перевірки статистичної значущості та побудови довірчих інтервалів оцінок параметрів регресії є Microsoft Excel. Його статистична функція ЛИНЕЙН дозволяє разом із МНК-оцінками параметрів регресії β_j також отримати їх стандартні помилки s_{β_j}. Інструмент «Регресія» надбудови «Аналіз даних» Microsoft Excel надає інформацію про стандартні помилки s_{β_j} параметрів, фактичні значення критеріїв Ст'юдента t_{β_j} та будує довірчі інтервали параметрів β_j. Критичне значення критерію Ст'юдента в Microsoft Excel можна визначити за допомогою статистичної функції СТЬЮДРАСПОБР. При її використанні в рядок «Вірогідність» її діалогового вікна записують не вірогідність p, а рівень значущості α.

8.2. Оцінка значущості моделі множинної регресії вцілому

Коефіцієнт детермінації – одна з найбільш ефективних оцінок адекватності регресійної лінійної моделі, міра якості рівняння регресії, характеристика його прогностичної сили [1]. Визначається як:

$$R^2 = 1 - \frac{(Y - X \cdot \beta)^T \cdot (Y - X \cdot \beta)}{(Y - \overline{Y})^T \cdot (Y - \overline{Y})}. \tag{8.7}$$

де $\overline{Y} = (\overline{y}, \overline{y}, \ldots, \overline{y})$.

Для моделі множинної лінійної регресії, побудованої за МНК коефіцієнт детермінації можна визначити возвівши множинний коефіцієнт кореляції в квадраті:

$$R^2 = r^2_{мн.} \tag{8.8}$$

Нагадаємо, що R^2 може змінюватись від 0 до 1 та характеризує частку варіації залежної змінної, обумовлену регресією або мінливістю незалежних змінних; чим ближче R^2 до одиниці, тим краще регресія описує залежність між незалежними і залежною змінними.

Разом з тим використання тільки одного коефіцієнта детермінації R^2 для вибору найкращого рівняння регресії може виявитися недостатнім. На практиці зустрічаються випадки, коли погано визначена модель регресії може дати порівняно високий коефіцієнт детермінації.

Недоліком коефіцієнтів детермінації R^2 є те, що він, загалом кажучи, збільшується при додаванні нових незалежних змінних, хоча це і не обов'язково означає покращення якості моделі множинної регресії. Тому краще використовувати *скоректований* (адаптований, поправлений) *коефіцієнт детермінації*:

$$R^2_{ск.} = 1 - \frac{n-1}{n-(m+1)} \cdot (1 - R^2). \tag{8.9}$$

Із (8.9) видно, що чим більше число m незалежних змінних, тим менше $R^2_{ск}$ в порівнянні з R^2. На відміну від R^2 скоректований коефіцієнт $R^2_{ск}$ може зменшуватися при введенні в модель нових незалежних змінних, що істотно не впливають на залежну змінну. Однак навіть збільшення скоректованого коефіцієнта детермінація при введенні в модель нової незалежної змінної не завжди означає, що вона покращилась.

Перевірка значущості рівняння множинної лінійної регресії здійснюється за F- критерієм Фішера, фактичне значення якого визначається як:

$$F_{факт.} = \frac{R^2 \cdot [n-(m+1)]}{(1-R^2)\cdot m}. \qquad (8.10)$$

Для заданого рівня значимості α, ступенів свободи k_1 й k_2 визначається критичне значення критерію Фішера. Причому, $k_1 = m$ і $k_2 = n-(m+1)$. При $F > F_{\alpha;k_1;k_2}$, рівняння регресії статистично значуще.

Статистична функція ЛИНЕЙН Microsoft Excel дозволяє отримати коефіцієнт детермінації R^2 та фактичне значення критерію Фішера $F_{факт.}$, інструмент «Регресія» надбудови «Аналіз даних» також надає значення скорегованого коефіцієнту детермінації $R^2_{ск}$. Критичне значення критерію Фішера $F_{\alpha;k_1;k_2}$ в Microsoft Excel можна визначити за допомогою статистичної функції FРАСПОБР. При її використанні в рядок «Вірогідність» її діалогового вікна записують не вірогідність p, а рівень значущості α.

8.3. Прогнозування на основі моделей множинної лінійної регресії

Як і моделі парної лінійної регресії, множинна лінійна регресія відноситься до так званих дескриптивних моделей, мета яких полягає в математичному описанні економічного явища та прогнозуванні, як правило короткостроковому, його розвитку. Оцінене за МНК рівняння регресії дозволяє описати економічне явище, зв'язки між характеристиками якого носять характер близький до лінійного.

Точковий прогноз вибіркової оцінки умовного математичного очікування залежної змінної може бути знайдений за рівнянням регресії:

$$M[y]_{X_0} = \beta_0 + \beta_1 \cdot x_{10} + \beta_2 \cdot x_{20} + \ldots + \beta_m \cdot x_{m0}. \qquad (8.11)$$

де X_0 – вектор прогнозованих значень кожної із незалежних змінних $X_0^T = (1, \quad x_{10}, \quad \ldots, \quad x_{m0})$.

Тоді p-відсотковий довірчий інтервал вибіркової оцінки умовного математичного очікування залежної змінної:

$$\left(M[y]_{X_0} - t_{p,f} \cdot s_{M[y]_{X_0}}\right) \le M[y]_{X_0} \le \left(M[y]_{X_0} + t_{p,f} \cdot s_{M[y]_{X_0}}\right); \qquad (8.12)$$

де $s_{M[y]_{X_0}}$ – стандартна помилка умовного математичного очікування $M[y]_{X_0}$.

$$s_{M[y]_{X_0}} = s \cdot \sqrt{X_0^T \cdot (X^T \cdot X)^{-1} \cdot X_0}\,. \qquad (8.13)$$

В деяких підручника наприклад [1] умовне математичне очікування ототожнюють із груповою середньою, відповідно рівняння (12.35) називають стандартною помилкою групової середньої.

Довірчий інтервал (8.12) враховує лише помилки та неточності побудови рівняння регресії, однак індивідуальні прогнозні значення залежної змінної $\hat{y}_0$ можуть відхилятися від рівняння регресії на величину випадкової помилки ε. Дисперсія цієї випадкової помилки ε визначається, як $\sigma_\varepsilon^2 = s^2$ та може бути оцінена за формулами (8.3) або (8.4). Таким чином стандартна помилка індивідуальних прогнозних значень залежної змінної $\hat{y}_0$ має буди більшою $s_{M[y]_{X_0}}$ на величину s та розраховуватися за формулою:

$$s_{\hat{y}_0} = s \cdot \sqrt{1 + X_0^T \cdot (X^T \cdot X)^{-1} \cdot X_0}\,. \qquad (8.14)$$

Тоді p-відсотковий довірчий інтервал для індивідуальних прогнозних значень залежної змінної:

$$\left(M[y]_{X_0} - t_{p,f} \cdot s_{\hat{y}_0}\right) \le \hat{y}_0 \le \left(M[y]_{X_0} + t_{p,f} \cdot s_{\hat{y}_0}\right). \qquad (8.15)$$

При прогнозуванні на основі регресійних моделей потрібно пам’ятати, що окрім похибок вихідних вибірок, побудови моделі і т.д., самі прогнозні значення незалежних змінних $X_0^T = (1, \; x_{10}, \; \ldots, \; x_{m0})$ можуть бути містити помилки.

При розгляді пункту 8.3 приймалося, що елементи вектору X_0 відомі точно, тобто ми мали справу із *безумовним прогнозуванням*. В [4] розглянуті основні принципи *умовного прогнозування* в регресійних моделях коли X_0 відомі нам лише приблизно.

При протіканні реальних економічних процесів на них впливають дуже багато чинників. Очевидно, що врахувати всіх їх при побудові моделі не можливо. Тому прогноз здійснений на базі регресійної моделі не можна вважати остаточною істиною. Він може виступати лише у якості допоміжної інформації, а управлінське рішення має приймати спеціаліст в тій чи іншій галузі економіки.

Запитання для самоперевірки

1. Описати алгоритм перевірки статистичної значущості параметрів множинної лінійної регресії за критерієм Ст'юдента.
2. Описати алгоритм побудови довірчих інтервалів для параметрів множинної лінійної регресії.
3. Комп’ютерні засоби оцінки статистичної значущості та побудови довірчих інтервалів для параметрів множинної лінійної регресії.
4. Коефіцієнт детермінації, як міра якості лінійної регресійної моделі.
5. Порівняти звичайний та скорегований коефіцієнт детермінації.
6. Причини, що призводять до необхідності розрахунку скорегованого коефіцієнту детермінації.
7. Перевірка статистичної значущості коефіцієнта детермінації.
8. Комп’ютерні засоби оцінки якості моделі множинної лінійної регресії.

Список літератури

1. Кремер Н.Ш., Путко Б.А. Эконометрика: Учебник для вузов / Под. ред. проф. Н.Ш. Кремера. – М.: ЮНИТИ-ДАНА, 2007. – 311 с.
2. Наконечний С.І., Терещенко Т.О., Романюк Т.П. Економетрія: Підручник. – К.: КНЕУ, 2005. – 520 с.
3. Эконометрика: Учебник / Под ред. И.И. Елисеевой. – М.: Финансы и статистика, 2004. – 344 с.
4. Магнус Я.Р., Катышев П.К., Пересецкий А.А. Эконометрика. Начальный курс: Учеб. – 6-е изд., перераб. и доп. – М.: Дело, 2004. – 576 с.
5. Введение в эконометрику / К. Доугерти ; пер. с англ. – М: ИНФРА-М, 1999. – 402 с.
6. Джонстон Дж. Эконометрические методы / Дж. Джонстон ; пер. с англ. – М.: Статистика, 1980. – 444 с.